

# Establishing consistency and improving uncertainty estimates of variational inference through M-estimation

Ted Westling\*

Department of Statistics, University of Washington  
and

Tyler H. McCormick

Department of Statistics, University of Washington

February 18, 2016

## Abstract

Variational inference (VI) is gaining popularity as a scalable estimation procedure for latent variable models. VI often empirically achieves similar predictive performance to slower, exact alternatives, but less is known about the viability of VI in contexts where parameter estimation and model interpretation are the primary goals. In this paper we connect VI for independent, identically distributed (IID) mixture models to M-estimation. We leverage extensive results about M-estimators from statistical theory to provide general conditions for consistency and asymptotic normality of VI point estimators. We also derive a "sandwich" asymptotic covariance matrix and a consistent estimator thereof. Our estimated covariance can be used to construct valid confidence regions and tests and is robust to model misspecification. We provide more specific conditions for the Gaussian Variational Approximation (GVA), which has been implemented in broad generality in the open-source software Stan. We conduct a thorough simulation study demonstrating our derived covariance matrix under correct and misspecified models and apply our methods to estimate a mixed effects logistic regression using data from the National Longitudinal Study of Adolescent Health.

*Keywords:* GLMM, mixture model, profile

---

\*The authors gratefully acknowledge Army Research Office grant 62389-CS-YIP.

# 1 Introduction

Computationally efficient estimation of a Euclidean parameter  $\theta$  in independent, identically distributed (IID) models with intractable likelihoods such as parametric mixture models is an important and active area of research. Methods based on the likelihood and inference based on the posterior distribution come with guarantees of asymptotic efficiency but are not computationally efficient in this setting. Variational inference (VI) is an increasingly popular method for approximating the MLE or posterior that does not require intensive integration. However, little is known about the properties of VI estimators of  $\theta$ . This paper provides regularity conditions for consistency and asymptotic normality of VI estimates of  $\theta$  that includes a model-robust formula for estimating uncertainty of variational estimators. The key insight of our work is connecting variational inference to  $M$ -estimation. Using a rich literature from statistics and econometrics, we leverage the properties of  $M$ -estimation to construct model-robust asymptotic covariance matrices for model parameters based only on derivatives of a function already calculated as part of the VI estimation process. Given that variational inference is primarily used when the amount of data is quite large, our asymptotic results are likely to frequently provide a good approximation.

We begin by defining parametric mixture models in both Bayesian and frequentist contexts and describing current tools for computation. We then connect VI to  $M$ -estimation and use this connection to provide conditions for consistency and asymptotic normality. Finally, we demonstrate the utility of our methods on simulated and real data.

Let  $X_1, \dots, X_n$  be  $p$ -variate data generated IID from an unknown distribution  $P_0$ . Let  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  be a model for the data where each  $P_\theta \in \mathcal{P}$  has an associated density

$$p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x, z) d\mu(z). \quad (1)$$

Here and throughout  $\theta$  is an unknown Euclidean parameter in  $\Theta \subset \mathbb{R}^d$ , and  $\mu$  is a dominating measure on  $\mathcal{Z} \subseteq \mathbb{R}^k$ . This is called a *parametric mixture model* or *latent variable model*, where  $(X_1, Z_1, \dots, X_n, Z_n)$  represent the full data, of which only  $(X_1, \dots, X_n)$  are observed, while  $Z_1, \dots, Z_n$  are latent or missing. In a Bayesian setting a prior distribution  $\pi(\cdot)$  is defined on the parameter vector  $\theta$ . If the model is correctly specified, then define  $\theta_0$  as the element of  $\Theta$  such that  $P_0 = P_{\theta_0}$ . Otherwise define  $\theta_0$  as the pseudo-true parameter maximizing

$\theta \mapsto P_0 \log p_\theta$ . In both cases assume  $\theta_0$  is the only such element of  $\Theta$ .

Parametric mixture models are widely used in many applied settings. For instance, random effects models are a type of mixture model used in regression analysis to represent unobserved heterogeneity across units in a population or correlation within repeated observations of a single unit (e.g. McCulloch and Neuhaus, 2001). Hierarchical models are an extension of random effects models to situations where units are part of a nested structure — e.g. students within classrooms within schools within districts — and the modeler would like to account for heterogeneity at each level (e.g. Gelman and Hill, 2006). Finite mixture models provide a model-based approach to clustering of continuous (e.g. Fraley and Raftery, 2002), categorical (e.g. Blei et al., 2003), and many other types of data (e.g. McLachlan and Peel, 2004). Mixed-membership models extend finite mixture models to allow units to belong to multiple clusters to varying extents (Airoldi et al., 2014).

When the integral in (1) has a closed form  $\theta$  can be estimated via *marginal* MLE or posterior approximation. Best Linear Unbiased Predictions (BLUP) can be used to obtain point estimates of the latent variables  $Z_{1:n}$  (Robinson, 2012) or EM can be used to simultaneously obtain the MLE and the conditional distribution of  $Z_{1:n}$  (Dempster et al., 1977).

Here we are concerned with the more difficult case when the integral in (1) is not closed-form, which is true in many models of interest (e.g. Generalized Linear Mixed Models). Marginal MLE of  $\theta$  is still possible by computing the likelihood, score, and information with numerical integration (Pinheiro and Chao, 2006) or by approximating the conditional distribution with Monte Carlo methods for use in an EM algorithm (McCulloch and Neuhaus, 2001). Bayesian estimation is also possible via approximation of the full posterior of  $\theta$  and the latent variables. These are all computationally expensive methods. Numerical integration is slow and noisy when the dimensionality of the latent variables is large or the function to be integrated is a product of many terms (see, e.g. McCulloch and Neuhaus (2001), Chapter 7). MCMC can suffer from slow convergence to the true posterior and high auto-correlation between successive samples, and these issues frequently become worse with more data and increasingly complex models (Gelman et al., 2014).

Given that obtaining the MLE or an MCMC approximation to the posterior can be computationally expensive, alternative methods that trade off some of the statistical guarantees of MLE and true posterior for improved computational scalability are desirable. This is es-

pecially true in the early stages of model-building when waiting hours or days for estimates is not practical, but may even be true for the final estimate if the model and data are especially onerous. Variational inference (VI) is one such approximate estimation method that emerged from the machine learning and statistical physics literatures and has proven to be both computationally and intuitively appealing (Wainwright and Jordan, 2008; Beal and Ghahramani, 2003). VI can be applied in likelihood contexts with variational EM (VEM) algorithms, and unlike other approximate solutions such as Penalized Quasi-Likelihood (Breslow et al., 1993), it can be applied to full Bayesian models as well with variational Bayes (VB) algorithms. The central idea in both cases is to replace the desired intractable conditional distribution  $p(z|x, \theta)$  or  $p(\theta, z|x)$  with the best possible approximation  $\hat{q}(z)$  in terms of Kullback-Leibler divergence in a class of distributions  $\mathcal{Q}$ . Since  $\mathcal{Q}$  is chosen by the modeler, it can be designed to make the inference problem easier – e.g. by allowing a coordinate ascent algorithm with closed-form updates or by reducing the complexity of the numerical integration. (Of course, the choice of  $\mathcal{Q}$  also influences the quality of VI estimate, as we discuss further below). Integrated nested Laplace approximation (INLA) is another approximate estimation procedure for latent variable models that avoids the need for complicated numerical integration or MCMC (Rue et al., 2009). However, INLA is only applicable in latent Gaussian models and hence less versatile than VI. For this reason, we focus on VI in this paper.

The strength of VI relative to exact ML and Bayesian estimation is its computational efficiency. While numerical integration and MCMC can suffer from convergence, speed, and scalability issues, VI often reaches at least a local maximum quickly and scales well with model dimensions  $d$  and  $k$  and sample size  $n$ . However, whereas numerical integration and MCMC are capable of producing arbitrarily good approximations of the true marginal MLE or posterior distribution respectively given enough time and computational resources, VI is not in all but the simplest cases.

This fundamental difference means the properties of VI estimates are not immediately clear but rather need to be justified. Given that VI is primarily used as a fast approximate estimation tool in predictive settings, the most common justification for its use is its empirical performance on predictive measures such as held-out log likelihood. However, in contexts where model parameter interpretation is paramount, such predictive measures are insufficient as they do not imply anything about parameter point or uncertainty estimates. VI can also

suffer from computational issues in the form of local maxima due to non-convexity of the criterion function, which we will not discuss here.

The limited existing research evaluating the properties of VI parameter estimates indicates that in many common models the estimates are consistent, but that the VB posterior uncertainty underestimates the true posterior or resampling uncertainty. VEM or VB algorithms have been shown to give consistent point estimates in finite Gaussian mixture models (Wang and Titterton, 2006), Markovian models (Hall et al., 2002), Poisson GLMMs (Hall et al., 2009, 2011; Ormerod and Wand, 2012), and the Stochastic Blockmodel (Bickel et al., 2013). However, uncertainty estimates from VB have been found to be too small (Wang and Titterton, 2004). For a specific example where the VB variance can be arbitrarily bad, see Appendix A of Rue et al. (2009). One recent method for estimating the covariance matrix from VB algorithms relies on Linear Response theory from statistical physics (Giordano et al., 2015). The methods we present here are derived from an entirely different framework specific to IID models and have a few distinct advantages over this recent work. First, the existing work assumes and requires that the VI estimator is consistent without providing a way to assess consistency, which we do. Neither do they present conditions under which their approximation holds nor robustness to model misspecification.

In this paper we will derive asymptotic properties of variational inference in general classes of models. Our central contribution will be to show in Sections 2 and 3 that under mild conditions both VEM and VB are a form of  $M$ -estimation, a connection that has not yet been made to our knowledge. We will use this connection to provide conditions for consistency of VI algorithms and to derive accurate model-robust asymptotic covariance matrices for model parameter estimates. In Section 4 we conduct a simulation study of our method applied to a specific mixture model, demonstrating its improvement over the VB posterior as well as its robustness to model misspecification. Finally, in Section 5 we apply our methods to derive more specific conditions for consistency and asymptotic normality of the Gaussian Variational Approximation (GVA), which has been implemented for a large class of parametric mixture models using Black Box Variational Inference (BBVI) in the open-source software Stan (Ranganath et al., 2014; Stan Development Team, 2015). We then apply our method to correct the covariance of a mixed effects logistic regression estimated with Stan using data from the National Longitudinal Study of Adolescent Health.

## 2 Variational EM asymptotics via $M$ -estimation

Our goal in this section is to demonstrate that variational EM point estimation of model parameters is a form of  $M$ -estimation and use this framework to provide conditions for consistency and asymptotic normality of the estimates. We begin by reviewing the variational EM approach to estimation of mixture models.

Variational EM requires specifying a *variational class*, also called a *variational family* of distributions  $\mathcal{Q}_n$  over the latent variables  $Z_{1:n}$  such that all distributions  $Q \in \mathcal{Q}_n$  are dominated by the true conditional distribution  $\pi(Z_{1:n}|X_{1:n}, \theta)$ . We will assume that the variational family is an independent product of the same class of distributions for each latent variable, i.e.  $\mathcal{Q}_n = \mathcal{Q}^n$  and  $q(Z_{1:n}) = \prod_i q_i(Z_i)$  for all  $q \in \mathcal{Q}_n$  and each  $q_i \in \mathcal{Q}$ . This is justified theoretically since for each fixed  $\theta$ , the true conditional distribution  $\pi(Z_{1:n}|X_{1:n}, \theta) = \prod_i \pi_i(Z_i|X_i, \theta)$  factors over the data as well.

Given  $\mathcal{Q}$ , define the *variational EM* estimate as

$$(\hat{\theta}_n, \hat{q}_{1:n}) \equiv \arg \max_{\theta \in \Theta, q_{1:n} \in \mathcal{Q}^n} \sum_{i=1}^n \mathbb{E}_{q_i} \left[ \log \frac{p(X_i, Z|\theta)}{q_i(Z)} \right]. \quad (2)$$

Let  $\mathcal{L}_n(\theta, q_{1:n}; X_{1:n})$  be the criterion function in (2).  $\mathcal{L}_n$  is called the *VEM ELBO*, where ELBO stands for Evidence Lower BOund, since it is a lower bound on  $\log p(X_{1:n}|\theta)$ , the log marginal likelihood or “evidence”.

Note that when  $\mathcal{Q}$  includes the true conditional distribution  $\pi(Z_i|X_i, \theta)$ ,  $\hat{q}_i = \pi(\cdot|X_i, \theta)$  and if this is true for all  $i$  then variational EM is equivalent to EM and  $\hat{\theta}_n$  is the true MLE. However, as discussed in the introduction, typically variational EM is used when the true conditional does not have a closed form, but  $\mathcal{Q}$  is taken to be a convenient parametric family nonetheless. In these cases  $\hat{q}_i$  is not the true conditional but rather the closest member of  $\mathcal{Q}$  to the true conditional (in KL divergence) and  $\hat{\theta}_n$  is not the MLE. Hence it is not clear whether  $\hat{\theta}_n$  is consistent or what its asymptotic distribution is.

We propose studying the properties of  $\hat{\theta}_n$  obtained by VEM through the lens of  $M$ -estimation.  $M$ -estimation is a general technique for constructing parameter estimates in parametric, semiparametric, and nonparametric models (see, e.g. Bickel et al. (1998)). The basic form of  $M$ -estimation for a parameter  $\theta$  with data  $X_{1:n}$  defines an estimator  $\tilde{\theta}_n$  for  $\theta_0$  as

the maximizer of the empirical average of a criterion function  $m$ :  $\tilde{\theta}_n = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^n m(\theta; X_i)$ . The criterion  $m$  is constructed such that the population criterion  $\mathbb{E}_{\theta_0}[m(\theta; X)]$  is maximized at  $\theta = \theta_0$ .

From Equation (2) we can see that  $\mathcal{L}_n(\theta, q_{1:n}; X_{1:n}) = \sum_{i=1}^n \ell(\theta, q_i; X_i)$  where  $\ell(\theta, q; x) = \mathbb{E}_q[\log \frac{p(x, z|\theta)}{q(z)}]$ . In this form the estimate doesn't fit in the  $M$ -estimation framework due to the dependence on  $q_i$ . In the statistical and econometric literature,  $q_i$  are known as *incidental* parameters specific to each data point, as opposed to  $\theta$ , which are the *structural* parameters shared across all data. However, by writing the optimization as a two-stage procedure, where first  $\mathcal{L}_n$  is optimized with respect to  $q_{1:n}$  for each fixed  $\theta$  and second this *profiled* ELBO is optimized with respect to  $\theta$ , we can view the estimation of  $\theta$  as an optimization over a parameter space with fixed dimension, hence enabling the use of  $M$ -estimation theory.

Formally, we have:

**Proposition 1.** *Let  $(\hat{\theta}_n, \hat{q}_{1:n})$  be the VEM estimate as defined in (2). Then*

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m(\theta; X_i)$$

$$\text{for } m(\theta; x) \equiv \sup_{q \in \mathcal{Q}} \ell(\theta, q; x) = \sup_{q \in \mathcal{Q}} \mathbb{E}_q \left[ \log \frac{p(x, Z|\theta)}{q(Z)} \right].$$

This approach is called *profiling* and is a common way of dealing with nuisance parameters in statistical estimation theory (Murphy and van der Vaart, 2000). However, we note that the profiling approach only allows us to demonstrate properties of the model parameters, not the latent variables, which is fine since we are concerned here with population-level model interpretation rather than unit-level interpretation or prediction.

The benefit of framing VEM as  $M$ -estimation is that it allows us to derive properties of  $\hat{\theta}_n$  in terms of the asymptotic criterion function  $\mathbb{E}_{P_0}[\sup_{q \in \mathcal{Q}} \ell(\theta, q; X)]$ , where  $P_0$  is the true distribution of  $x$  (i.e.  $P_0 = P_{\theta_0}$  if the model is correctly specified). In particular we can determine to what  $\hat{\theta}_n$  is converging. We have the following result:

**Theorem 2** (Consistency of VEM). *Suppose the function  $\mathbb{E}_{P_0}[\sup_{q \in \mathcal{Q}} \ell(\theta, q; X)]$  attains a finite global maximum at  $\bar{\theta}$  and conditions (A1)-(A3) hold. Then  $\hat{\theta}_n \rightarrow_{P_0} \bar{\theta}$ .*

The proofs of all results are presented in Appendix B.

If the model is correctly specified so that  $P_0 = P_{\theta_0}$  for some  $\theta_0 \in \Theta$ , then variational EM is consistent if  $\bar{\theta} = \theta_0$ . This result provides a way to validate the consistency of VEM algorithms. The key condition that must be satisfied is that  $\mathbb{E}_{P_0}[\sup_{q \in \mathcal{Q}} \ell(\theta, q; X)]$  is uniquely maximized at the (pseudo-)true parameter  $\theta_0$ .

When the model is correctly specified we can provide a somewhat more intuitive form for this key condition. Let  $\Pi_{x,\theta}$  be the conditional distribution of  $\mathbf{Z}$  given  $x$  and  $\theta$ . When for all  $\theta$  and  $x$  the supremum over  $\mathcal{Q}$  is achieved at a unique element of  $\mathcal{Q}$ , we can define  $Q_{\theta,x}^* = \arg \max_{q \in \mathcal{Q}} \ell(\theta, q; x)$ . Then  $m(\theta; x) = \sup_{q \in \mathcal{Q}} \ell(\theta, q; x) = \ell(\theta, q_{\theta,x}^*; x)$ . Using this, a more transparent and perhaps convenient form of the key condition for consistency of VEM is:

**Proposition 3.**  $\mathbb{E}_{\theta_0}[\sup_{q \in \mathcal{Q}} \ell(\theta, q; X)]$  is maximized at  $\theta_0$  if and only if

$$\theta \mapsto D_{KL}(P_{\theta_0} \| P_{\theta}) + \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta,X}^* \| \Pi_{\theta,X})] \quad (3)$$

is uniquely minimized at  $\theta = \theta_0$ .

In many types of variational inference the variational class  $\mathcal{Q}$  is either assumed to be parametrized by a finite-dimensional Euclidean parameter  $\psi \in \Psi$  or it can be proven that the optimal variational distribution lies in such a finite-dimensional subclass (as in mean-field algorithms for exponential families). Throughout the rest of the paper we will assume this is the case to avoid a discussion of functional derivatives. We will call  $\psi$  the *variational parameters* and will write  $\ell(\theta, \psi; x)$  to refer to  $\ell(\theta, q_{\psi}; x)$  and  $\psi_{\theta,x}^* = \arg \max_{\psi \in \Psi} \ell(\theta, \psi; x)$ .

Before moving on to asymptotic normality of VEM estimators, it is worth mentioning a related perspective based on estimation equations. When the criterion function  $\ell$  is differentiable in  $\theta$  and  $\psi$  it is sometimes more natural to view the algorithm as finding a root of the gradient  $\nabla \ell$ , which is known as an estimating equation or  $Z$ -estimator (or sometimes, confusingly, also as an  $M$ -estimator). Then under regularity conditions (Godambe, 1991) the variational estimate  $\hat{\theta}_n$  converges to the  $\bar{\theta}$ , the unique element of  $\Theta$  satisfying  $\mathbb{E}_{\theta_0}[\nabla_{\theta} m(\bar{\theta}; X)] = \mathbf{0}$ . As we demonstrate in the proof of Theorem 4 in the appendix,  $\nabla_{\theta} m(\theta; x) = (\nabla_{\theta} \ell)(\theta, \psi^*(\theta; x); x)$ . Hence evaluating the consistency of  $\hat{\theta}_n$  using the estimating equations framework amounts to evaluating  $\mathbb{E}_{\theta_0}[(\nabla_{\theta} \ell)(\theta, \psi^*(\theta; X); X)] = \mathbf{0}$ . The estimating equations approach is in some cases more mathematically convenient, but it is



only useful when the population estimating equation has a single root, which can be limiting when the profiled criterion has many local modes (see, e.g. Example 5.50 from van der Vaart (2000) concerning the score equation for a Cauchy location parameter).

The  $M$ -estimation framework also provides results about asymptotic normality of VEM estimators, and as we will show, the asymptotic covariance matrix can be estimated consistently whether or not the criterion function  $m$  is known explicitly. Conditions on  $m$  can then be replaced by conditions on  $\ell$  and  $\psi_{\theta,x}^*$ . In particular, we have the following result:

**Theorem 4** (Asymptotic Normality of VEM). *Suppose  $\hat{\theta}_n \rightarrow_{P_0} \theta_0$  and conditions (A4)-(A10) hold. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\mathbf{H}(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} m(\theta_0; X_i) + o_P(1) \quad (4)$$

where

$$\mathbf{H}(\theta) = \mathbb{E}_{P_0}[D_{\theta\theta}^2 m(\theta, X)]. \quad (5)$$

Thus,  $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N_d(0, \mathbf{V}(\theta_0)^{-1})$  where

$$\mathbf{V}(\theta) = \mathbf{H}(\theta)^{-1} \mathbb{E}_{P_0} \left[ (\nabla_{\theta} m(\theta_0; X_i)) (\nabla_{\theta} m(\theta_0; X_i))^T \right] \mathbf{H}(\theta_0)^{-1}. \quad (6)$$

Furthermore,  $\hat{\mathbf{V}}_n(\hat{\theta}_n) \rightarrow_P \mathbf{V}(\theta_0)$  where  $\hat{\mathbf{V}}_n(\theta) = \hat{\mathbf{H}}_n(\theta)^{-1} \hat{\mathbf{B}}_n(\theta) \hat{\mathbf{H}}_n(\theta)^{-1}$  for

$$\hat{\mathbf{H}}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left[ D_{\theta\theta}^2 \ell(\theta, \hat{\psi}_i; X_i) - D_{\theta\psi}^2 \ell(\theta, \hat{\psi}_i; X_i) \left( D_{\psi\psi}^2 \ell(\theta, \hat{\psi}_i; X_i) \right)^{-1} D_{\theta\psi}^2 \ell(\theta, \hat{\psi}_i; X_i)^T \right] \quad (7)$$

and

$$\hat{\mathbf{B}}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left( \nabla_{\theta} \ell(\theta, \hat{\psi}_i; X_i) \right) \left( \nabla_{\theta} \ell(\theta, \hat{\psi}_i; X_i) \right)^T. \quad (8)$$

The asymptotic covariance in equation (6) is known as a *sandwich* covariance due to the way  $\mathbf{H}$  is on both sides of the expression.

The estimate of the asymptotic variance provided in equations (7) and (8) only relies on knowing derivatives of  $\ell$  and the maximizing values  $\hat{\psi}_{1:n}$ . Hence these formulas provide a way to construct correct and model-robust asymptotic covariance matrices for model parameters based only on derivatives of a function already calculated as part of the VI estimation process.

In some cases  $m(\theta; x)$  can be derived explicitly, in which case the empirical average of the second derivative matrix  $\frac{1}{n} \sum_i D_{\theta\theta}^2 m(\hat{\theta}_n; X_i)$  can be used as a simpler estimate of  $\hat{\mathbf{H}}_n(\theta_0)$ . Also note that the function being averaged in equation (7) is the Schur complement of the lower right block in the hessian matrix of  $\ell$  with respect to  $(\theta, \psi)$  evaluated at  $X_i$ .

### 3 Variational Bayes

In a fully Bayesian context with a prior distribution  $\Pi_0$  over  $\theta$ , VB algorithms extend VEM by specifying a class of variational posteriors  $\mathcal{S}$  over  $\theta$ . (In theory the variational class may include distributions with dependence between  $Z$  and  $\theta$ , but in practice the two classes are almost always assumed independent.) The variational criterion function then becomes

$$\tilde{\mathcal{L}}_n(s, q; X_{1:n}) = \mathbb{E}_{s,q} \left[ \log \frac{p(X_{1:n}, Z_{1:n}, \theta)}{q(Z_{1:n})s(\theta)} \right] = \mathbb{E}_s [\mathcal{L}_n(\theta, q; X_{1:n})] - D_{KL}(S \parallel \Pi_0). \quad (9)$$

Note that in VB the criterion function minimizes the KL divergence between  $s(\theta) \prod_i q_i(Z_i)$  for  $s \in \mathcal{S}, q_i \in \mathcal{Q}$  and the true posterior distribution  $\Pi_n(\theta, Z_1, \dots, Z_n)$ .

A clear trend in the literature on VB inference is that even when the VB posterior is consistent, it frequently underestimates the true posterior variance (Wang and Titterton, 2004; Rue et al., 2009). This phenomenon can be explained intuitively using the KL divergence between multivariate normal distributions. Under regularity conditions, the posterior distribution of model parameters  $\Pi_n(\theta|X_{1:n})$  looks approximately  $N_D(\theta_0, \Sigma)$  as  $n$  grows (where  $\Sigma$  implicitly depends on  $n$  because  $\theta$  has not been appropriately rescaled). Often the variational distribution is asymptotically normal as well (Wang and Titterton, 2010). However, if the variational class of distributions over model parameters only includes factored distributions, then the variational distribution can only be approaching independent normal distributions. A few lines of algebra show that  $D_{KL}(N(\cdot|\mu, \text{diag}(\sigma_1^2, \dots, \sigma_D^2)) \parallel N(\cdot|\theta_0, \Sigma))$ , the KL divergence between such an independent multivariate normal variational distribution and a general multivariate normal, is minimized when  $\mu = \theta_0$  and  $\sigma_k^2 = 1/(\Sigma^{-1})_{kk}$ . However, using Schur complements we can see that

$$\sigma_k^2 = \frac{1}{(\Sigma^{-1})_{kk}} = \Sigma_{kk} - \Sigma_{k\cdot}(\Sigma_{-kk})^{-1}\Sigma_{\cdot k}^T. \quad (10)$$

where  $\Sigma_{k\cdot}$  is the  $k$ th row of  $\Sigma$  omitting  $\Sigma_{kk}$  and  $\Sigma_{-kk}$  is the minor of  $\Sigma$  removing the  $k$ th row and  $k$ th column. Assuming  $\Sigma$  is positive definite,  $\Sigma_{-kk}^{-1}$  is positive definite as well and hence  $\Sigma_{k\cdot}\Sigma_{-kk}^{-1}\Sigma_{k\cdot}^T \geq 0$  with equality if and only if  $\Sigma_{k\cdot} = \mathbf{0}$ . Hence  $\sigma_k^2$ , the marginal variational posterior variance of  $\theta_k$ , is  $\leq \Sigma_{kk}$ , the true marginal posterior variance, with equality if and only if  $\theta_k$  is not correlated in the posterior with any of the other model parameters! The above says roughly that we should expect the VB posterior to underestimate the marginal uncertainty of any model parameter or latent variable that is asymptotically correlated with other model parameters or latent variables.

Given that the posterior from VB is not generally a reliable assessment of the uncertainty in  $\hat{\theta}_n$ , we would like to get a better covariance estimate for constructing intervals and regions. Since VB effectively puts a prior on the VEM estimates, as long as the prior is fixed relative to sample size and has support in a neighborhood of the true parameter, we can expect that VB point estimates (e.g. the VB posterior mean or median) will be within  $o_p(1/\sqrt{n})$  of the VEM estimates. Hence when the VEM estimates converge at a  $\sqrt{n}$  rate, the limiting behavior of the VB point estimates will be the same as that of the VEM estimates.

We provide intuition here for how the same analysis used for establishing consistency of VEM estimates could be used for VB estimates. Assume the variational distribution for  $\theta$  is parametrized by  $(\mu, \omega) \in \Theta \times \Omega$ , where  $\theta$  is the VB point estimate and  $\omega$  controls the dispersion in the sense that  $\theta_n \rightarrow_{L_1} \mu$  as  $\omega_n \rightarrow \omega_0$ , a limit point of  $\Omega$ . This ensures that the variational family includes point masses at each  $\theta \in \Theta$  as  $L_1$  limit points. Then defining  $\tilde{\Omega} = \Omega \cup \{\omega_0\}$  and  $\tilde{\ell}(\mu, \omega, q; x) = \mathbb{E}_{\mu, \omega}[\ell(\theta, q; x)]$ , we can see that the VB estimate  $(\mu_n^*, \omega_n^*, q_{1:n}^*)$  approximately maximizes  $\sum_i \tilde{\ell}(\mu, \omega, q; X_i)$  (the KL divergence between the variational distribution and prior for model parameters plays no role in the limit).

Furthermore, as with VEM we can define  $\tilde{m}(\mu, \omega; x) = \sup_q \tilde{\ell}(\mu, \omega, q; x)$  so that  $(\mu_n^*, \omega_n^*)$  can be seen to approximately optimize the profiled criterion function  $\frac{1}{n} \sum_i \tilde{m}(\mu, \omega; X_i)$  over  $\Theta \times \Omega$ . Then under appropriate smoothness conditions on  $\ell$  and the same key population maximization condition on  $m$  we would have that  $(\mu_n^*, \omega_n^*) \rightarrow_{P_0} (\bar{\theta}, \omega_0)$ ; i.e. the VB posterior  $s_{\mu_n^*, \omega_n^*}$  over  $\theta$  converges to a point mass at  $\bar{\theta}$ .

The asymptotic distribution of the VB posterior could perhaps be obtained by again applying M-estimation techniques to  $\tilde{m}$ , but since  $\omega_n^*$  is not converging to an interior point of  $\Omega$ , alternative techniques would be needed than those used in the proof of Theorem 4.

However, for the purposes of deriving the marginal limit distribution of  $\mu_n^*$  and constructing asymptotic intervals and tests of the model parameter, it would be sufficient to show that  $\frac{1}{n} \sum_{i=1}^n m(\hat{\mu}_n^*; X_i) \geq \frac{1}{n} \sum_{i=1}^n m(\hat{\theta}_n; X_i) - o_P(1/n)$ . This would imply that the asymptotic covariance matrix of the VB point estimates is the same as that derived for the VEM estimates.

## 4 Case study: Exponential mixture model

As a first example we will consider the following model: IID pairs  $(X_i, Y_i)$  are observed for  $i = 1, \dots, n$  with  $X_i|\theta \sim \exp(\theta)$ ,  $Y_i|\theta, Z_i \sim \exp(\theta Z_i)$  conditionally independent given  $Z_i$ , where  $Z_i$  is a latent variable with distribution  $Z_i|\lambda \sim \exp(\lambda)$ .  $(\theta, \lambda) \in \mathbb{R}^{+2}$  is the parameter vector. Marginally  $Y_i|\theta, \lambda$  has density  $p(y|\theta, \lambda) = (1/(\lambda/\theta)) \frac{1}{(y/(\lambda/\theta)+1)^2}$ , which is a Pareto type II distribution (also known as a Lomax distribution) with location  $\mu = 0$ , scale  $\lambda/\theta$  and shape  $\alpha = 1$ . Furthermore  $Z_i|Y_i, X_i, \theta, \lambda$  has a posterior  $\text{Gamma}(2, \theta y + \lambda)$  distribution. Hence in this example both marginal maximum likelihood over  $(\theta, \lambda)$  or an EM algorithm provide asymptotically efficient point estimates for  $(\theta, \lambda)$  with the usual inverse information asymptotic covariance matrix. However, for the purpose of demonstrating our results we can derive variational algorithms for this model and investigate their properties.

### 4.1 Consistency and asymptotic distribution

Consider a VB algorithm with a fully factored variational class  $\mathcal{Q}_n = \{Q : q(\theta, \lambda, Z_{1:n}) = q(\theta)q(\lambda) \prod_i q(Z_i)\}$ . A variational class like this which includes all fully factored distributions is called a *mean-field* variational class. When the model is a conditionally conjugate exponential family and the mean-field variational class is used, there are well-known explicit formulas for a computationally efficient coordinate ascent algorithm (Wainwright and Jordan, 2008). In this case, with Gamma prior distributions  $\theta \sim \text{Gamma}(\alpha_{0\theta}, \beta_{0\theta})$  and  $\lambda \sim \text{Gamma}(\alpha_{0\lambda}, \beta_{0\lambda})$  the mean-field coordinate ascent algorithm has the usual closed-form updates because the model is composed of conditionally conjugate exponential families, which we derive in full in the supplementary material.

Since the mean-field variational class of distributions over  $Z_{1:n}$  includes the Gamma posterior, VEM is equivalent to EM and hence the VEM estimates are equivalent to the marginal

MLE. This can also be seen by plugging in the variational updates for the distribution of  $Z_i$  in to the ELBO. Hence  $\hat{\theta}_n$  and  $\hat{\lambda}_n$  obtained via mean-field VEM will be exactly equal to the MLE and thus under correct model specification will be consistent and asymptotically normally distributed with covariance equal to the negative inverse Fisher information matrix, which turns out to be  $\begin{pmatrix} \theta_0^2 & \theta_0 \lambda_0 \\ \theta_0 \lambda_0 & 4\lambda_0^2 \end{pmatrix}$ .

With fixed conjugate Gamma priors on  $\theta$  and  $\lambda$ , the point estimates from VB converge to these same values with the same asymptotic covariance, and we can demonstrate this formally by examining the full VB ELBO. However, the VB posterior will underestimate the true marginal variances. The variational posterior distribution over  $(\theta, \lambda)$  is a product of two gamma distributions with mean converging to  $(\theta_0, \lambda_0)$  and  $n\text{Var}(\hat{\theta}_n - \theta_0) \rightarrow \theta_0^2/2$ ,  $n\text{Var}(\hat{\lambda}_n - \lambda_0) \rightarrow \lambda_0^2$  (with covariance fixed to 0 because we are using a mean-field algorithm). For  $\theta$ , the VB marginal variance underestimates the true asymptotic variance by a factor of two, while for  $\lambda$  it underestimates by a factor of four. Note that this is actually worse than if we had performed variational inference on the marginal distribution of  $(\theta, \lambda)$  using independent distributions, in which case the marginal variance would have been underestimated by a factor of  $\frac{4}{3}$  for both  $\theta$  and  $\lambda$ , which can be seen by using (10).

In the supplementary material we demonstrate that a variational algorithm that uses a parametric variational class that is smaller than the mean-field class, which is called *fixed-form* variational inference, is also consistent with the same asymptotic distribution. As we discuss below, this is the type of approximation made by Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2015).

## 4.2 Case Study I: Exponential mixture model

To empirically validate the theoretical results presented here, we simulated 1000 samples with  $\theta_0 = 1$  and  $\lambda_0 = 1$  for values of  $n$  from 10 to 10,000 and fit the variational algorithm to the model. We tried other values of  $\theta_0$  and  $\lambda_0$  and found that the only difference was the rate at which the true coverage approaches the nominal coverage, which makes sense given that normal approximations are affected by skewness. Constructing the covariance and intervals on the  $\log \theta$  and  $\log \lambda$  scales would reduce the skewness and improve the approximation for small sample sizes.

We initialized the mean-field algorithm using simple moment estimators:  $\theta_{\text{start}} = 1/\bar{\mathbf{X}}$ ,

$\lambda_{\text{start}} = \theta_{\text{start}} \mathbf{Y}_{50}$ , and  $(Z_i)_{\text{start}} = 2/(\theta_{\text{start}} Y_i + \lambda_{\text{start}})$ . We found that the initialization was very important for the asymptotics of VB estimates and hence performance of various intervals. The asymptotics discussed here concern the global maximizer of the ELBO, but the VB ELBO has multiple local maxima, and the initialization affects which local maxima is found. As a result, initializing too far from the true values results in an estimator with slightly larger than optimal asymptotic variance, and hence intervals which are slightly narrow (i.e. around 90% rather than the nominal 95%). On the other hand, initializing at the true values (which of course is impossible in practice) results in a local maximum closer to the truth than the global maximizer, and hence an estimator with asymptotic variance slightly smaller than optimal and intervals which are too wide.

We constructed 95% confidence intervals/regions for  $\theta_0$ ,  $\lambda_0$ , and  $(\theta_0, \lambda_0)$  jointly using 1) the variational posterior, 2) the inverse Fisher information matrix of the marginal model, and 3) the sandwich covariance matrix. Although the variational updates of local parameters would allow us to derive the second derivative matrix explicitly, since this is not always possible we used the general formula from Theorem 4 to construct the covariance matrix instead. For each of the three intervals/regions and three covariances we estimated 1) the resampling (frequentist) coverage and 2) the posterior (Bayesian) coverage. The resampling coverage is the proportion of constructed regions containing the true parameter value over all simulations. The posterior coverage is the posterior probability contained in the interval. The true marginal posterior distribution of  $(\theta, \lambda)$  was approximated using Stan (Stan Development Team, 2015).

The results of the simulations under correct model specification are shown in Figure 1. The variational posterior was consistent, but its intervals were too narrow (the 95% regions using the VB posterior are anti-conservative). This is in agreement with previous findings in the literature, and as discussed above is due to the underestimation of the marginal variance when KL-approximating a correlated multivariate distribution with independent distributions. Both the information and sandwich had correct coverage for both the marginal and joint parameters by about  $n = 50$ . Also note that the frequentist and Bayesian empirical coverage rates are very similar.

Figure 2 shows the variance (rescaled by a factor of  $n$ ) as estimated by VB, inverse information matrix, sandwich, and MCMC samples. First note that all but VB are converging

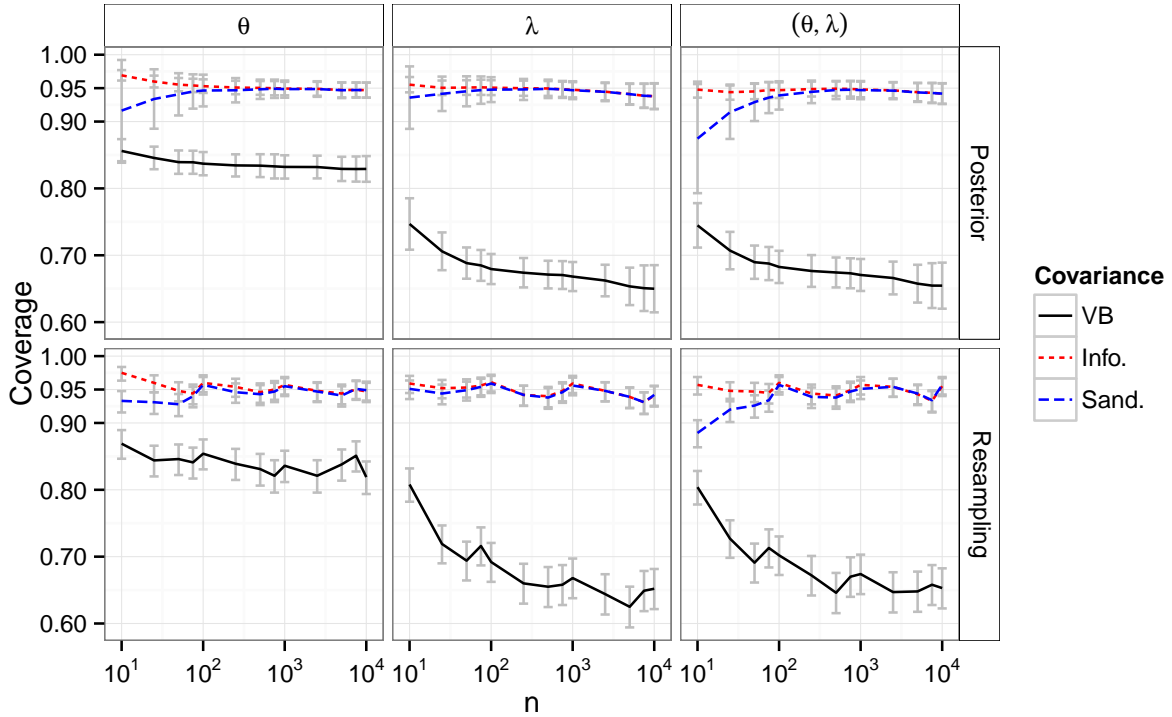


Figure 1: Coverage results of 95% confidence intervals/regions from the exponential mixture model under correct model specification. Line type indicates covariance used to construct the interval. Panels show marginal coverage of  $\theta$ ,  $\lambda$  and joint coverage of  $(\theta, \lambda)$ . Top row shows mean and  $\pm 1$  standard deviation of the posterior (Bayesian) coverage. Bottom row shows resampling (frequentist) coverage with 95% binomial CIs of the estimated coverage.

to the same value as  $n$  grows, which is exactly equal to the corresponding element of the negative inverse information matrix. The marginal variance of  $\theta$  and  $\lambda$  from VB is converging to a smaller value, which for  $\theta$  is  $1/2$  and for  $\lambda$  is  $1$ . Since  $\theta_0 = \lambda_0 = 1$ , these are the exact values for the asymptotic variance of VB as predicted by the theory. Of course the covariance is zero for VB for all  $n$  by design.

While the inverse information matrix performs well under correct model specification in this example, we don't expect this to always be the case. It happens to be correct here because it turns out the variational algorithm is equivalent to maximum likelihood, and we include it in the simulations to verify this fact, but that certainly isn't always true. In any practical example truly requiring the use of variational inference (which this example does not), the variational estimators are almost certainly not equivalent to the (marginal) MLE, and hence the inverse information would not be an appropriate estimate of the asymptotic

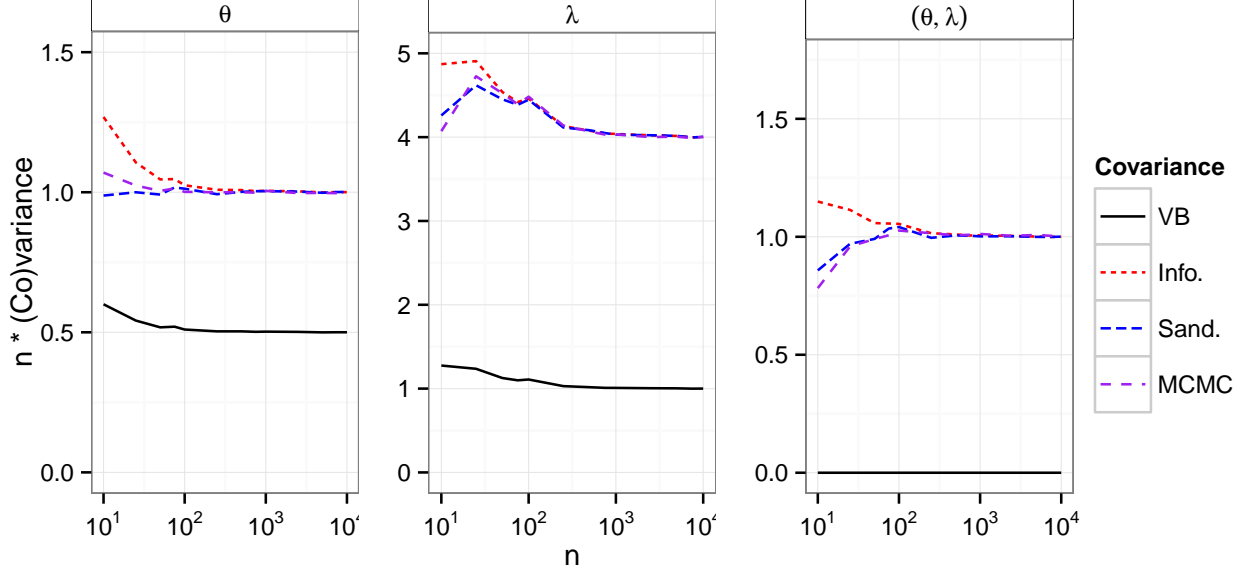


Figure 2: Rescaled (co)variances of model parameter estimates as a function of  $n$ . Line type represent estimates of the (co)variance. Panels from left to right are the marginal variances of  $\theta$  and  $\lambda$  and the covariance of  $\theta$  and  $\lambda$ .

covariance of the model parameters. Additionally, the Fisher information is only easy to use in this case since the marginal likelihood can be calculated in closed form. If that were not the case then calculating the information would require numerical integration.

Next we simulated data under model misspecification. Instead of drawing  $X_i \sim \text{Exp}(\theta)$ , we drew it from  $\text{Gamma}(3, 3\theta)$ . In this case since the VEM ELBO only depends on the distribution of  $X$  through its mean and the mean of this misspecified distribution is still  $1/\theta$ , the pseudo-true parameter is equal to the true parameter and the algorithm will still be consistent. However, the point estimates will have a different covariance than when the model was correctly specified, so in this case we don't expect the inverse Fisher information to provide correct uncertainty estimates. We also do not expect the true posterior distribution to be an accurate assessment of the uncertainty in  $\hat{\theta}$ .

We conducted the same simulation study for this model-misspecification as above. The results are shown in Figure 3. In this case, both VB and the inverse information intervals are conservative for  $\theta$  in terms of resampling coverage. Only the sandwich achieves the nominal coverage. The posterior coverage results indicate that the inverse information is a very good approximation for the true posterior covariance, which in turn indicates that the posterior covariance is too small relative to the resampling covariance. Although the



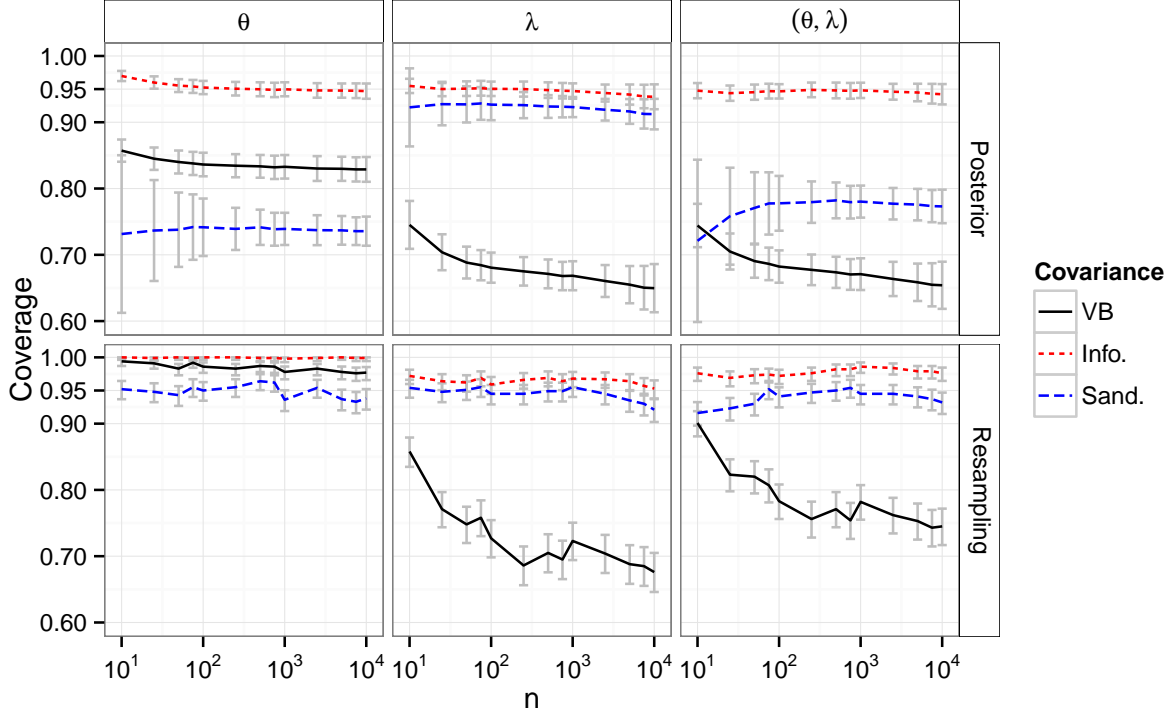


Figure 3: Coverage results of 95% confidence intervals/regions from the exponential mixture model under model misspecification where  $X \sim \text{Gamma}(3, 3\theta)$ .

sandwich covariance is smaller than the posterior covariance, it more closely resembles the resampling covariance and provides better frequentist coverage.

## 5 Case Study II: Mixed effects logistic regression

Next we apply our method to logistic regression with random intercepts. For subjects  $i = 1, \dots, n$  and observations  $k = 1, \dots, m_i$  (the number  $m_i$  of which may vary by subject) let  $Y_{ik}|\gamma_i, \mathbf{X}_{ik}$  be IID Bernoulli( $\text{logit}^{-1}(\gamma_i + \mathbf{X}_{ik}\boldsymbol{\beta})$ ) and  $\gamma_i$  be IID  $N(0, \sigma^2)$ .  $\mathbf{X}_i$  are observed covariates and  $\gamma_i$  is the random intercept for subject  $i$ . The parameter vector is  $(\boldsymbol{\beta}, \sigma)$ , where  $\sigma$  is the standard deviation of the random intercepts.

Since this model is not a conditionally conjugate exponential family, there is no closed-form mean-field coordinate ascent algorithm. Recently a method called Automatic Differentiation Variational Inference (ADVI) that makes VI possible for such models has been implemented in the open-source statistical estimation software Stan (Kucukelbir et al., 2015;

Stan Development Team, 2015). This represents a significant advancement in the potential application of variational inference, as it makes estimation using variational inference as easy as specifying the model in Stan’s modeling language. In this section we will demonstrate how our methods can be used in conjunction with ADVI and Stan to assess consistency of these types of VI algorithms and to correct the estimated covariance matrix. We then estimate a logistic regression with random intercepts using Stan and data from the National Longitudinal Study of Adolescent Health.

## 5.1 Gaussian Variational Approximation

ADVI uses an independent Gaussian Variational Approximation for latent variables and model parameters. As we have discussed, our approach only depends on the variational approximation for the latent variables, only using the point estimates for the model parameters (i.e. the VEM approximation), so for convenience we will omit discussion of the variational distribution for the model parameters. Suppose the latent variables are  $k$ -dimensional, i.e.  $\mathbf{Z} = (Z_1, \dots, Z_K)^T$ , so that  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iK})$ . When the model is defined with the latent variables on a domain other than the entire real line, ADVI first transforms each  $Z_k$  to  $\mathbb{R}$ . For notational simplicity we will assume  $\mathbf{Z}$  has already been transformed (so that the Jacobian of the transformation is absorbed in  $p$ ). Then ADVI takes the variational approximation to be an independent Gaussian over each  $Z_k$ :  $q(\mathbf{Z}|\mathbf{m}, \mathbf{s}) = \prod_k dN(Z_k; m_k, s_k^2)$ . The ELBO works out to (ignoring constant terms):

$$\mathcal{L}_n(\boldsymbol{\theta}, \mathbf{m}_{1:n}, \mathbf{s}_{1:n}; \mathbf{X}_{1:n}) = \sum_{i=1}^n \left[ \mathbb{E}_{\mathbf{m}_i, \mathbf{s}_i} [\log p_{\boldsymbol{\theta}}(\mathbf{X}_i, \mathbf{Z}_i)] + \sum_k \log s_{ik} \right]. \quad (11)$$

The expectation is approximated using Monte Carlo integration or  $K$  univariate numerical integrations, which removes the restriction that models produce closed-form variational expectations. Using our notation, the asymptotic properties of the variational algorithm depends on  $\ell(\boldsymbol{\theta}, \mathbf{m}, \mathbf{s}; \mathbf{x}) = \mathbb{E}_{\mathbf{m}, \mathbf{s}} [\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z})] + \sum_k \log s_k$ .

We first revisit the conditions for consistency stated in Section 2. Recall the key that  $\mathbb{E}_{P_0}[\ell(\boldsymbol{\theta}, \psi_{\boldsymbol{\theta}, X}^*; X)]$  be maximized uniquely at  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . Finding the function in the expectation can be onerous when  $\psi_{\boldsymbol{\theta}, x}^*$  is not known in closed form. In the case of GVA we can find

Method	$\beta_0$	$\beta_1$	$\sigma$
ADVI	0.831	0.860	0.411
Sand.	0.938	0.955	0.959

Table 1: True coverage rates of nominal 95% CIs constructed with the ADVI variational posterior and the sandwich correction in a simulation study with  $n = 100$  subjects of logistic regression with random intercept.

implicit equations satisfied by  $\psi_{\theta,x}^*$ , thus easing the demonstration of consistency in any particular case. Furthermore, since the expectations are all with respect to univariate Gaussian distributions, we can apply Stein’s lemma to obtain multiple equivalent forms for each of the derivatives, which we list in Appendix A. Hence assuming  $\ell$  is uniquely maximized for each  $\theta$  and  $\mathbf{x}$ , we have that these expressions evaluated at  $\mathbf{m}^*$  and  $\mathbf{s}^*$  are all zero.

Next we present similar formulas for the “sandwich” asymptotic covariance for GVA. The asymptotic covariance requires computing the partial derivatives of  $\ell$  with respect to  $\theta$  and the variational parameters, which can be found by pushing the derivatives inside the expectation when necessary. We present these formulas in Appendix A. As in ADVI, the expectations in these derivatives are approximated using Monte Carlo integration by sampling from normal distributions. The converged values  $\mathbf{m} = \mathbf{m}^*$  and  $\mathbf{s} = \mathbf{s}^*$  will be plugged in to obtain the asymptotic covariance.

Table 1 shows the results of a simulation study comparing the true coverage of nominal 95% intervals obtained by ADVI and our sandwich formula for a random-intercept logistic regression with  $n = 100$ ,  $m_i \sim 3 + \text{Poisson}(10)$ ,  $\beta = (-1, 2)$  and  $\sigma = 0.5$ . The ADVI intervals underestimate while the sandwich intervals correctly assess the the true uncertainty, as expected.

In the next section we explore using ADVI and our sandwich covariance together to estimate a mixed logistic regression on real data from the National Longitudinal Study of Adolescent Health.

## 5.2 Application: National Longitudinal Study of Adolescent Health

We applied our method to estimate a logistic regression with random intercepts to data from the National Longitudinal Study of Adolescent Health (AddHealth) (Harris, 2009). We followed Shin et al. (2009) in constructing a dataset using waves I and III of the study

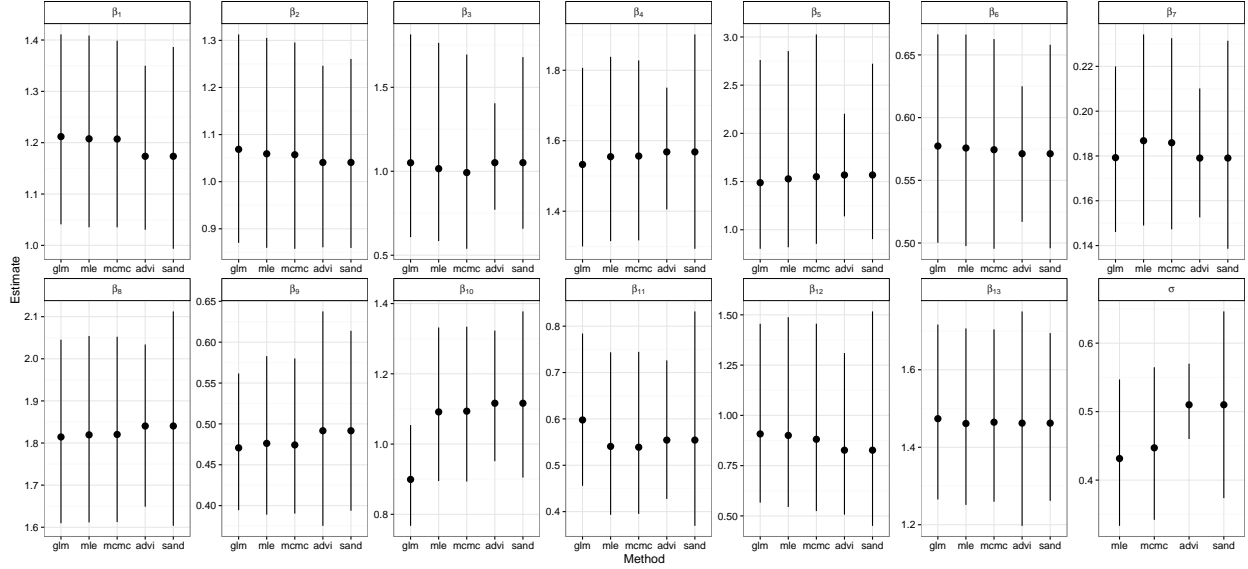


Figure 4: Top: AddHealth data exponentiated coefficients and 95% confidence/coverage intervals, corresponding the same covariates in the same order as those listed in Shin et al. (2009) Table 3, as well as the standard deviation  $\sigma$  of school random effects.

to analyze the relationship between binge drinking in adolescents (a binary outcome) and childhood maltreatment (a set of binary predictors) adjusting for a number of potential confounders in 12,966 adolescents. We were not able to match their dataset or results exactly due to an incomplete listing of exclusion restrictions, but we have constructed an analysis that matches their general setup. Shin et al. (2009) conducted a logistic regression to determine this relationship, but given that subjects are clustered in schools, it is reasonable to expect that which school a subject attends is related to whether they binge drink. To explore this, we adapted the analysis of Shin et al. (2009) to include a random intercept for each high school.

We estimated the model in four ways. First, we repeated the analysis of Shin et al. (2009) by estimating a logistic regression without random effects. We then estimated the model with high school random intercepts via marginal MLE with a Gauss-Hermite approximation to the log-likelihood, via an MCMC approximation to the posterior distribution, and via ADVI with Stan. Furthermore, for the ADVI estimate, we have both the ADVI variance estimate of the coefficient as well as the sandwich variance estimate.

Figure 4 shows the five estimates and 95% CIs for the thirteen covariates presented in Table 2 of Shin et al. (2009) as well as the standard deviation of school random effects.

The thirteen covariates are (all indicator variables): neglect only, physical abuse only, sexual abuse, neglect and physical abuse, all type of abuse, age 15-17, age 12-14, male, race black, ethnicity hispanic, race asian, race other, parent alcoholism. Also included in the regression but not reported: mother works outside the home, father works outside the home.

The ADVI point estimate is similar to (and always well within the 95% CI of) the MLE and MCMC point estimates, but the 95% CI from ADVI is frequently either much smaller or much larger than those of MLE and MCMC. The sandwich covariance provides uncertainty estimates for the ADVI point estimate are more in agreement with the remaining methods. On average the sandwich 95% CIs for the fixed effects contained 95.0% of the corresponding marginal posterior distribution, compared to just 80.9% for the ADVI posteriors (the minimums were 89.1% and 30.7% respectively). The joint multivariate normal sandwich confidence region for the fixed effects contained 80.0% of the joint posterior mass, while joint ADVI region contained just 18.0%. Hence ADVI point estimates with the corrected sandwich posterior is a viable and computationally efficient alternative to MLE or MCMC.

We also note that it does not appear that including school-level random effects changes the substantive conclusions of the original logistic regression.

As with sandwich covariance approximations in other regression contexts, whether the covariates are fixed or random matters if the model is misspecified (Buja et al., 2015). Here we have implicitly assumed random covariates.

## 6 Conclusion

We have demonstrated that Variational Inference applied to IID mixture models is a form of  $M$ -estimation, thus connecting it with a rich statistical theory. We have used this theory to derive methods for establishing consistency and asymptotic normality of VI point estimators of model parameters, thus making it possible to consider VI as an estimation strategy in settings where parameter interpretation and inference are paramount. The asymptotic covariance matrix can be estimated consistently, even under model misspecification, using derivatives of the variational criterion function.

Our case study demonstrates how our method can be used to establish consistency of variational algorithms as well as construct the asymptotic covariance matrix. Our simulations

demonstrate that this covariance has good properties from both a frequentist and Bayesian point of view and beats alternative methods under model misspecification. Our application demonstrates that our method can be used in conjunction with the open-source software Stan to produce point estimates and confidence regions that are comparable to slower, exact methods such as MLE and MCMC with no work on the user’s part beyond specifying the model. This suggests variational inference should be taken seriously as an estimation method for large-scale or complicated IID mixture models when the primary goal is hypothesis testing and effect estimation.

## A Gaussian Variational Approximation derivatives

First derivatives of the ELBO: Let  $\boldsymbol{\psi} = (\mathbf{m}, \mathbf{s})^T$  be the vector of GVA variational parameters and define  $W_k = \frac{Z_k - m_k}{s_k}$ . Assuming the necessary differentiability and pushing derivatives in to the expectation, we obtain

$$\frac{\partial \ell}{\partial m_k} = \frac{1}{s_k} \mathbb{E}_{\mathbf{m}, \mathbf{s}} [W_k \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z})] = \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ \frac{\partial}{\partial Z_k} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] \quad (12)$$

$$\frac{\partial \ell}{\partial s_k} = \frac{1}{s_k} \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ (W_k^2 - 1) \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] + \frac{1}{s_k} = \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ W_k \frac{\partial}{\partial Z_k} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] + \frac{1}{s_k} \quad (13)$$

$$= s_k \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ \frac{\partial^2}{\partial Z_k^2} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] + \frac{1}{s_k} \quad (14)$$

For the second derivative with respect to  $\boldsymbol{\theta}$  and the mixed derivatives we simply push the derivative with respect to  $\boldsymbol{\theta}$  inside the expectations above, giving e.g.  $\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2} = \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right]$  and  $\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial m_k} = \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ \frac{\partial}{\partial \boldsymbol{\theta} \partial Z_k} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right]$ .

The second derivatives with respect to the variational parameters admit multiple equivalent representations, of which we only list a few.

$$\frac{\partial^2 \ell}{\partial m_k \partial m_j} = \frac{1}{s_k s_j} \mathbb{E}_{\mathbf{m}, \mathbf{s}} [(W_k W_j - \mathbf{1}_{k=j}) \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z})] = \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ \frac{\partial^2}{\partial Z_k \partial Z_j} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] \quad (15)$$

$$\frac{\partial^2 \ell}{\partial m_k \partial s_j} = \frac{1}{s_k s_j} \mathbb{E}_{\mathbf{m}, \mathbf{s}} [W_k (W_j^2 - 1 - 2 \cdot \mathbf{1}_{k=j}) \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z})] \quad (16)$$

$$= \frac{1}{s_k} \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ (W_k W_j - \mathbf{1}_{k=j}) \frac{\partial}{\partial Z_j} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] \quad (17)$$

$$= \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ W_j \frac{\partial^2}{\partial Z_k \partial Z_j} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right]. \quad (18)$$

$$\frac{\partial^2 \ell}{\partial s_j \partial s_k} = \frac{1}{s_k s_j} \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ ((W_k^2 - 1)(W_j^2 - 1) - \mathbf{1}_{j=k}(3W_k^2 - 1)) \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] - \mathbf{1}_{j=k} \frac{1}{s_k^2} \quad (19)$$

$$= \frac{1}{s_k} \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ W_j (W_k^2 - 1 - \mathbf{1}_{k=j}) \frac{\partial}{\partial Z_j} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] - \mathbf{1}_{j=k} \frac{1}{s_k^2} \quad (20)$$

$$= \mathbb{E}_{\mathbf{m}, \mathbf{s}} \left[ W_k W_j \frac{\partial^2}{\partial Z_k \partial Z_j} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{Z}) \right] - \mathbf{1}_{j=k} \frac{1}{s_k^2}. \quad (21)$$

## B Proof of theorems

Recall that  $P_0$  is the true distribution,  $\mathcal{Q}$  is the variational family of distributions over the latent variable  $Z$  and

$$\ell(\theta, q; x) = \mathbb{E}_q \left[ \log \frac{p(x, Z|\theta)}{q(Z)} \right]$$

is one term in the VEM ELBO.

We start with a list of assumptions we will need for consistency of VEM:

**(A1)** The map  $\theta \mapsto \sup_{q \in \mathcal{Q}} \ell(\theta, q; x)$  is upper-semicontinuous a.s.- $P_0$ .

**(A2)** There exists a  $d > 0$  such that for all  $\delta < d$  and  $\eta \in \Theta$  the map

$$x \mapsto \sup_{\substack{\theta \in B_\delta(\eta) \\ q \in \mathcal{Q}}} \ell(\theta, q; x)$$

is measurable and

$$\mathbb{E}_{P_0} \sup_{\substack{\theta \in B_\delta(\eta) \\ q \in \mathcal{Q}}} \ell(\theta, q; X) < \infty.$$

**(A3)** There exists a compact set  $K \subset \Theta$  such that  $P_0(\hat{\theta}_n \in K) \rightarrow 1$ .

Theorem 2 follows from these conditions.

*Proof of Theorem 2.* Defining  $m(\theta; x) = \sup_{q \in \mathcal{Q}} \ell(\theta, q; x)$ , the requirements of Theorem 5.14 of van der Vaart (2000) are satisfied. Hence for all  $\epsilon > 0$ ,  $P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) \rightarrow 0$ . Since  $P_0(\hat{\theta}_n \in K) \rightarrow 1$  by assumption,

$$\begin{aligned} P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon) &\leq P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) + P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K^c) \\ &\leq P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) + P_0(\hat{\theta}_n \in K^c) \rightarrow 0. \end{aligned}$$

□

Next we prove Proposition 3 relating consistency of VEM to KL divergence.

*Proof of Proposition 3.* Note that

$$\ell(\theta, q; X) = \log p_\theta(X) - D_{KL}(Q \| \Pi_{X, \theta}). \quad (22)$$

Hence  $\mathbb{E}_{\theta_0}[\ell(\theta, q_{\theta, x}^*; X)] < \mathbb{E}_{\theta_0}[\ell(\theta_0, q_{\theta_0, x}^*; X)]$  if and only if

$$\mathbb{E}_{\theta_0}[\log p_\theta(X)] - \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta, X}^* \| \Pi_{\theta, X})] < \mathbb{E}_{\theta_0}[\log p_{\theta_0}(X)] - \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta_0, X}^* \| \Pi_{\theta_0, X})] \quad (23)$$

which can be reexpressed as

$$\begin{aligned} \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta_0, X}^* \| \Pi_{\theta_0, X})] &< \mathbb{E}_{\theta_0}[\log p_{\theta_0}(X) - \log p_\theta(X)] + \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta, X}^* \| \Pi_{\theta, X})] \\ &= D_{KL}(P_{\theta_0} \| P_\theta) + \mathbb{E}_{\theta_0}[D_{KL}(Q_{\theta, X}^* \| \Pi_{\theta, X})]. \end{aligned}$$

□

We next lay out sufficient conditions for asymptotic normality.

- (A4) For all  $\theta$  and  $P_0$ -a.e.  $x$ ,  $\ell(\theta, \psi; x)$  is uniquely maximized at  $\psi^*(\theta; x)$  which is an element of  $\Psi$ , an open subset of  $\mathbb{R}^d$ .
- (A5) For all  $\theta$ ,  $(\psi, x) \mapsto \ell(\theta, \psi; x)$  is a measurable function on the product measure space  $\Psi \times \mathcal{X}$ , where  $\Psi$  is equipped with Borel measure.
- (A6) For all  $\theta$ ,  $x \mapsto \psi^*(\theta; x)$  is a measurable function of  $x$ .
- (A7)  $\psi^*$  is twice continuously differentiable in a neighborhood of  $\theta_0$  for  $P_0$ -a.e.  $x$ .
- (A8)  $\ell$  is twice continuously differentiable in a neighborhood of  $\theta_0$  and  $\psi^*(\theta_0; x)$  for  $P_0$ -a.e.  $x$ .
- (A9) There exist  $r > 0$ ,  $s(x) > 0$ ,  $b_1(x)$  and  $b_2(x)$  such that

$$(a) \text{ For all } x \in \mathcal{X} \text{ and } \theta \in \mathcal{B}_r(\theta_0), \psi^*(\theta; x) \in \mathcal{B}_{s(x)}(\psi^*(\theta_0; x))$$



(b) For all  $x \in \mathcal{X}$ ,  $\theta_1, \theta_2 \in \mathcal{B}_r(\theta_0)$  and  $\psi_1, \psi_2 \in \mathcal{B}_{s(x)}(\psi^*(\theta_0; x))$ ,

$$|\ell(\theta_1, \psi_1; x) - \ell(\theta_2, \psi_2; x)| \leq b_1(x)(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|).$$

(c) For all  $\theta_1, \theta_2 \in \mathcal{B}_r(\theta_0)$ ,  $\|\psi^*(\theta_1; x) - \psi^*(\theta_2; x)\| \leq b_2(x)\|\theta_1 - \theta_2\|$ .

(d)  $b_1$  and  $b_1 b_2 \in L_2(P_0)$ .

**(A10)**  $|D_\theta^2 \ell(\theta, \psi^*(\theta; x); x)| \leq \kappa(x)$  for all  $\theta$  in a neighborhood of  $\theta_0$  and  $P_0$ -a.e.  $x$  for an integrable function  $\kappa$ .

With these conditions we prove Theorem 4.

*Proof of Theorem 4.* The main asymptotic linearity statement of Equation 4 will be proved using van der Vaart (2000) Theorem 5.23. We need to validate the following conditions to apply the result: 1.  $m(\theta; x)$  is measurable as a function of  $x$  for all  $\theta \in \Theta$ ; 2.  $m(\theta; x)$  is differentiable at  $\theta_0$  for  $P_0$ -a.e.  $x$ ; 3. there exists a measurable function  $b \in L_2(P_0)$  and an  $r > 0$  such that for all  $\theta_1, \theta_2 \in \mathcal{B}_r(\theta_0)$ ,  $|m(\theta_1; x) - m(\theta_2; x)| \leq b(x)\|\theta_1 - \theta_2\|$ ; 4. The function  $m(\theta) = \mathbb{E}_{P_0}[m(\theta; X)]$  is maximized at  $\theta = \theta_0$  and admits a second-order Taylor expansion at  $\theta_0$ ; 5.  $\frac{1}{n} \sum_i m(\hat{\theta}_n; X_i) \geq \sup_{\theta \in \Theta} \frac{1}{n} \sum_i m(\hat{\theta}; X_i) - o_P(1)$ . We will demonstrate that these conditions follow from assumptions (A4)-(A10).

For the first condition, the measurability of  $x \mapsto m(\theta; x)$  is guaranteed by (A5) and (A6) plus the fact that compositions of measurable functions are measurable.

Differentiability of  $m$  at  $\theta_0$  is implied by conditions (A7) and (A8) together with the multivariate chain rule. We have  $(\nabla_\theta m)(\theta_0; x) = (\nabla_\theta \ell)(\theta_0, \psi^*(\theta_0; x); x) + (D_\theta \psi^*)(\theta_0; x)^T (\nabla_\psi \ell)(\theta_0, \psi^*(\theta_0; x); x)$ . Since  $\psi \mapsto \ell(\theta_0, \psi; x)$  is maximized at  $\psi^*(\theta_0; x)$ , which is in the interior of  $\Psi$ , and  $\ell$  is differentiable in  $\psi$  at  $\theta_0, \psi^*(\theta_0; x)$  for  $P_0$ -a.e.  $x$ ,  $(\nabla_\psi \ell)(\theta_0, \psi^*(\theta_0; x); x) = 0$  a.s.  $P_0$ .

For the third condition we apply (A9). Let  $\theta_1, \theta_2 \in \mathcal{B}_r(\theta_0)$ . Then for each  $x$   $\psi^*(\theta_1; x), \psi^*(\theta_2; x) \in \mathcal{B}_{s(x)}(\psi^*(\theta_0; x))$ . Hence

$$\begin{aligned} |m(\theta_1; x) - m(\theta_2; x)| &= |\ell(\gamma^*(\theta_1; x); x) - \ell(\gamma^*(\theta_2; x); x)| \leq b_1(x) (\|\theta_1 - \theta_2\| + \|\psi^*(\theta_1; x) - \psi^*(\theta_2; x)\|) \\ &\leq b_1(x) (\|\theta_1 - \theta_2\| + b_2(x)\|\theta_1 - \theta_2\|) = b_1(x)(1 + b_2(x))\|\theta_1 - \theta_2\|. \end{aligned}$$

Since by assumption  $b_1, b_1 b_2 \in L_2(P_0)$ , the third condition is satisfied with  $b = b_1(1 + b_2)$ .

Assumptions (A7), (A8) and (A10) imply that the map  $\theta \mapsto \mathbb{E}_{P_0}[\ell(\theta, \psi^*(\theta; x); x)] = \mathbb{E}_{P_0}[m(\theta; x)]$  is twice continuously differentiable in a neighborhood of  $\theta_0$  and hence possesses a second-order Taylor expansion, thus satisfying the fourth condition above. Furthermore, these assumptions justify differentiation under the integral. We can thus derive the second derivative matrix of  $m(\theta; x)$  at  $\theta_0$  as follows:

$$D_{\theta\theta}^2 m(\theta; x) = D_\theta(\nabla_\theta m(\theta; x)) = D_\theta(\nabla_\theta \ell)(\theta_0, \psi^*(\theta_0; x); x) \quad (24)$$

$$= (D_{\theta\theta}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x) + (D_{\theta\psi}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x)(D_\theta \psi^*)(\theta_0; x) \quad (25)$$

By definition  $\psi^*(\theta_0; x)$  solves  $(\nabla_\psi \ell)(\theta_0, \psi^*(\theta_0; x); x) = 0$ . Differentiating with respect to  $\theta$  gives

$$0 = (D_{\theta\psi}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x) + (D_\theta \psi^*)(\theta_0; x)(D_{\psi\psi}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x) \quad (26)$$

$$(D_\theta \psi^*)(\theta_0; x) = -(D_{\theta\psi}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x)(D_{\psi\psi}^2 \ell)(\theta_0, \psi^*(\theta_0; x); x)^{-1}. \quad (27)$$

Solving for  $(D_\theta \psi^*)(\theta_0; x)$  and substituting this back in to (25) gives

$$D_{\theta\theta}^2 m(\theta; x) = (D_{\theta\theta}^2 \ell - (D_{\theta\psi}^2 \ell)(D_{\psi\psi}^2 \ell)^{-1}(D_{\theta\psi}^2 \ell)^T)(\theta_0, \psi^*(\theta_0; x); x). \quad (28)$$

All the derivatives are now in terms of  $\ell$ , a known function, and evaluated at  $\theta_0, \psi^*(\theta_0; x)$ .

Finally, condition five is satisfied since  $\hat{\theta}_n$  maximizes  $\frac{1}{n} \sum_i m(\hat{\theta}; X_i)$  by definition. This establishes the asymptotic linearity and normality statements of the theorem.

For the consistency of the estimator  $\hat{\mathbf{V}}_n(\hat{\theta}_n)$  for the true asymptotic covariance  $\mathbf{V}(\theta_0)$ , apply the WLLN and continuous mapping theorem to each component matrix.

□

## References

- Airoldi, E. M., D. M. Blei, E. A. Erosheva, and S. E. Fienberg (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- Beal, M. J. and Z. Ghahramani (2003). The Variational Bayesian EM Algorithm for Incom-

- plete Data : with Application to Scoring Graphical Model Structures. *Bayesian Statistics* 7.
- Bickel, P., D. Choi, X. Chang, and H. Zhang (2013, August). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics* 41(4), 1922–1943.
- Bickel, P. J., C. A. Klaassen, Y. Ritov, J. A. Wellner, et al. (1998). Efficient and adaptive estimation for semiparametric models.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993–1022.
- Breslow, N. E., N. E. Breslow, D. G. Clayton, and D. G. Clayton (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88(421), 9–25.
- Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang (2015). Models as approximations – a conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science* 1460.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (Methodological)* 39(1), 1–38.
- Fraley, C. and A. E. Raftery (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Taylor & Francis.
- Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Giordano, R., T. Broderick, and M. Jordan (2015). Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes.

- Godambe, V. P. (1991). *Estimating Functions*. Clarendon Press.
- Hall, P., K. Humphreys, and D. M. Titterton (2002, August). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3), 549–564.
- Hall, P., J. T. Ormerod, and M. Wand (2009). Theory of Gaussian variational approximation for a Poisson mixed model.
- Hall, P., T. Pham, M. Wand, and S. Wang (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *The Annals of Statistics* (January).
- Harris, K. M. (2009). The national longitudinal study of adolescent to adult health (add health), waves i & ii, 1994–1996; wave iii, 2001–2002; wave iv, 2007–2009 [machine-readable data file and documentation].
- Kucukelbir, A., R. Ranganath, A. Gelman, and D. M. Blei (2015). Automatic Variational Inference in Stan.
- McCulloch, C. E. and J. M. Neuhaus (2001). *Generalized Linear Mixed Models*. Wiley Online Library.
- McLachlan, G. J. and D. Peel (2004). *Finite mixture models*. John Wiley & Sons, Inc.
- Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association* 95(450), 449–465.
- Ormerod, J. T. and M. P. Wand (2012, January). Gaussian Variational Approximate Inference for Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics* 21(1), 2–17.
- Pinheiro, J. C. and E. C. Chao (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models.
- Ranganath, R., S. Gerrish, and D. M. Blei (2014). Black Box Variational Inference. In *AISTATS*, Volume 33.

- Robinson, G. K. (2012). That BLUP Is a Good Thing : The Estimation of Random Effects. *Statistical Science* 6(1), 15–32.
- Rue, H. v., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* 71(2), 319–392.
- Shin, S. H., E. M. Edwards, and T. Heeren (2009). Child abuse and neglect: relations to adolescent binge drinking in the national longitudinal study of adolescent health (addhealth) study. *Addictive behaviors* 34(3), 277–280.
- Stan Development Team (2015). CmdStan: the command-line interface to Stan, Version 2.8.0.
- van der Vaart, A. W. (2000). *Asymptotic statistics* (3 ed.). Cambridge University Press.
- Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1(1-2), 1—305.
- Wang, B. and D. M. Titterton (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations.
- Wang, B. and D. M. Titterton (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* (3), 625–650.
- Wang, B. and D. M. Titterton (2010). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 1–14.